CLARIN
A European Research Infrastructure
Newsletter

Number 2, 2008, July

TERENA – CLARIN Collaboration

Peter Wittenburg, Daan Broeder, Dieter Van Uytvanck

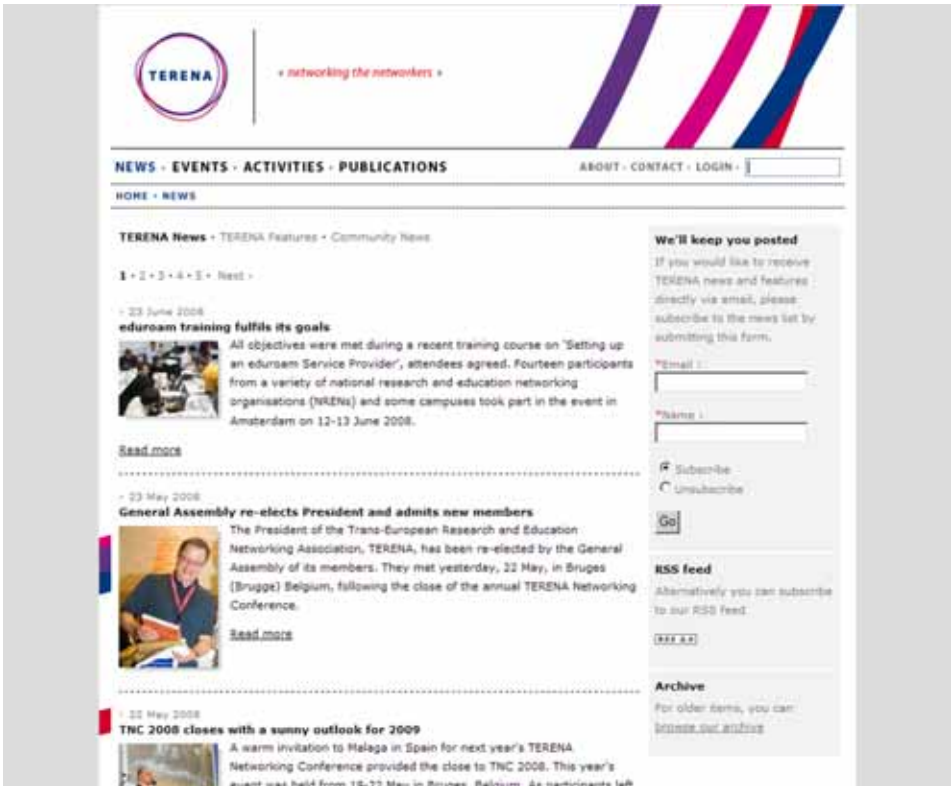
The Trans-European Research and Education Networking Association (TERENA, http://www.terena.org/) was set up to offer a forum to collaborate, innovate and share knowledge in order to foster the development of Internet technology, infrastructure and services to be used by the research and education community. One of the tasks of TERENA is to harmonize between the different national approaches in the area of distributed authentication and authorization. Different attribute sets to categorize and describe different users are being used and, when similar, sometimes they are not used the same way. All these differences hamper cross-national federations, and there is a need for harmonization. This is why TERENA set up, for example, the Schema Harmonization Committee (SCHAC, http://www.terena.org/activities/tf-emc2/schac.html) to: (1) build an internal kernel from existing local attributes and agree on syntax and semantics and let the kernel evolve via a collaborative approach; (2) define a method to accept new attributes/classes; (3) to promote the schemas within the NRENs (National Research and Education Network) constituency, making the results part of the local schemas; (4) promote the schemas in other fields, such as GEANT2 (European data communication network), digital libraries, EUNIS (European University Information Systems), Internet2 (US research network initiative), etc.

In July a first official meeting was organized at the TERENA office in Amsterdam bringing experts from TERENA and CLARIN together to discuss opportunities of a closer collaboration.

CLARIN wants to establish a federation of language resources and technologies covering institutes from many countries in Europe and come to agreements with the emerging national identity federations that hopefully will cover researchers from an increasing number of European academic institutions. The core of such federations are trust agreements where, in particular, issues of user management and the exchange of user attributes need to be specified.

The understanding that CLARIN needs to interact not only with the national identity providers, but in particular with the European initiative that wants to harmonize the various attempts led us to meet the secretary general, Karel Vietsch, and Licia Florio who is one of TERENA's technical experts.

Due to CLARIN's European scope, TERENA is a natural and expected ally. CLARIN seems to be one of the first initiatives that comes to TERENA with a large user community and a distributed service provider scenario. Therefore, it was agreed that CLARIN will explain its goals and wishes at one of the coming expert meetings, and that TERENA and CLARIN will keep in further close contact.



The home page of TERENA at http://www.terena.org

Editors' Foreword

Marko Tadić
& Dan Cristea
CLARIN Newsletter editors

Dear readers,
for many people who attended LREC this year, the magnificent Marrakesh is still extremely alive in memories: with its narrow and crowded streets full of colours and smells, with the snake charmers performing their skills in the open, with the shop owners almost grabbing you to visit their incredible bazaar-like exhibitions, with the bargaining habits reminding that we are already in the middle of Orient although at the Greenwich longitude, with the arabesque of sculptured wood and stone heavily decorating the Muslim palaces, with the romantic riads, the authentic old-style hotels built around inner patios where the tea arrives in the glass after executing a skilful and spectacular vault in the air.

It was there that we decided to distribute a historical, if you allow us to prefigure, first issue of our CLARIN Newsletter, as we wanted to exploit the advantage of having there more than a thousand of participants working in Language Resources and Technologies for putting it directly in their hands.

We are glad now to bring to your attention our second issue. You will notice that it respects the same main schema, that we tried to impose as a constant for the CLARIN Newsletter, mainly including minimally a parallel view of developers and consumers, a presentation of major events related to CLARIN, and reports describing recent developments in the CLARIN member states.

We have invited Peter Wittenburg to open the issue with a note on a recently set up collaboration between TERENA, another important collaborating project that builds research and educational infrastructures, and CLARIN. The reason why we have chosen this as the cover story is that we wanted thus to stress the importance of the interactivity with major pan-European initiatives intended to develop technological infrastructures that will be of help to researchers in the social sciences and humanities.

Next to this page you can read in an article by Maria Gavrilidou and Stelios Piperidis about the efforts that are being performed presently in Greece to preserve the cultural heritage and the role that the language resources and technologies play there.

The next page, bring the parallel views of consumer and the developer. This time Željko Hodonj from the Croatian News Agency has the contribution which describes how this company plans to use LR&T to further speed up their work-flow and also to make new information-broker products. The developer's view is presented professor Bojana Dalbelo Bašić from Zagreb University.

On page 5 Thierry Declerck discusses the topic which brings about the essential issue of standards in LR&T that will be of crucial importance to the CLARIN community.

We have selected for review in the middle pages three events that took place recently in Europe. LREC – the Language Resources and Evaluation Conference, where at least two meetings with great relevance for CLARIN took place. They are described by Steven Krawer. Then, the Digital Humanities Conference in Oulu, Finland is reviewed by Martin Wynne, and the ESF workshop on the role of humanities in CEE countries in Sofia, Bulgaria, presented by Marko Tadić.

Peter Wittenburg and Tamás Váradi are interviewed by us in an attempt to shed more light on the lively issue of the role of commercial companies in LR&T research in Europe, a topic which they addressed in an article in the previous issue of this newsletter. We felt that we should devote more space to the question of whetheher we are really in competition with the Internet and software giants.

The following articles describe two CLARIN national projects: the German one, as authored by Erhard Hinrichs, Peter Wittenburg, Alexander Geyken, Lothar Lemnitzer and Andreas Witt, and the Danish one, as authored by Hanne Fersøe. Enjoy your reading! C

Call for contributions

Dear readers
of the CLARIN Newsletter,
if you have ideas, thoughts, comments, additions, corrections, arguments, questions etc. which are connected to the CLARIN project, even remotely, please feel free to send them to us as your contribution at newsletter@clarin.eu.

List of national correspondents

Austria
Gerhard Budin

Belgium – Flanders
Inneke Schuurman

Bulgaria
Svetla Koeva

Croatia
Marko Tadić

Czech Republic
Karel Pala

Denmark
Hanne Fersøe

ELRE/ELDA
Bente Maegaard

UK
Martin Wynne

Estonia
Tiit Roosmaa

Finland
Kimmo Koskenniemi

France
William Del Mancino
Bertrand Gaiffe

Germany
Lothar Lemnitzer

Greece
Maria Gavrilidou
Stelios Piperidis

Hungary
Tamás Váradi

Italy
Valeria Quochi

Latvia
Andrejs Vasiljevs

Malta
Mike Rosner

Netherlands
Peter Wittenburg

Norway
Koenraad De Smedt

Portugal
Antonio Branco

Poland
Maciej Piasecki

Romania
Dan Cristea
Dan Tufiș

Spain
Nuria Bel

Sweden
Sven Strömqvist

Language technology in the Greek cultural heritage sector

Maria Gavrilidou, Stelios Piperidis
ILSP – Athena RC

In Greece the largest programme related to the Social Sciences and the Humanities is the *Operational Programme Information Society,* in the framework of which, more than 140 projects are currently being completed. These projects aim at the digitisation, documentation and promotion of the Greek cultural heritage, including all aspects of culture and civilisation. In this sense, museum objects, textual as well as audiovisual archival items, collections and archives of great historical value, theatre archives, folklore museums' material, music and painting collections, ecclesiastical art collections, costume collections etc. have been digitised, documented and made available to the public. The goal of this huge (for the country) endeavour was the preservation of the cultural heritage with the aid of information technologies, but also the exploitation of the cultural material, the development of related digital by-products and services, and last but not least, raising awareness about digital content among the citizens of the country. To this Call many cultural material holders responded, such as museums, libraries, archives, cultural organisations, etc.

Reccomendations and best practices

The *use of standards and intelligent techniques* at all levels of development has been explicitly stressed throughout the Programme. In the Preparatory Phase of the Programme several *studies* were completed which elaborated *recommendations* as regards *standards and best practices* to be used. The studies concerned: technologies for the digitisation of 2-D material (text, image), 3-D material (objects, monuments, archaeological sites), video and moving image, sound and music; technologies, standards and metadata for the interoperability, documentation, re-usability and management of digital content; technologies for the protection and management of IPR issues, and finally, for the development of portals and websites of the organisations, including recommendations and standards for multilingual content authoring and retrieval.

Focus on LT

Not unexpectedly, the studies concerning textual archives and associated metadata focused on *Language Technology* methodologies, standards, tools, applications and integrated systems, as regards all stages of creation, processing and maintenance of the digital content. Consequently, NLP and LRs know-how has been sought after by the cultural content holders, and it has been successfully deployed in the framework of this programme.

Most of these cultural organisations have sought support in *indexing their digitised primary or metadata material* in an as intelligent manner as possible. For instance, shallow semantic information extractors have been used like term and keyword extractors, named entity recognisers, etc., to automatically generate meaningful index terms. Such extracted indexical data have in many cases been structured in *thesauri* following ISO standards 2788 and 5964 for monolingual and multilingual thesauri. Furthermore, and in an attempt to provide unified, single-entry access to topically similar or related collections, *meta-thesaurus* structures have been implemented for those collections for which indexing had been performed on the basis of specific thesauri.

Multilingual access to the data was another key issue in this framework: following recommendations, the material (basically the documentation material, but in certain cases also primary data) had to be available in many languages. For the translation of the material, *translation memories* and *multilingual terminologies* have been used, where they already existed, and created, when needed. For the translation aids, the use of TMX and TBX was enforced, catering for compatibility with standards and ensuring exchange and reuse.

Interoperability

As for *interoperability* with similar efforts in the cultural domain, the recommended model was the *CIDOC Conceptual Reference Model* (CRM), which provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. Aiming to serve as a guide for good practice of conceptual modelling, it is now official standard ISO 21127:2006 "intended to promote a shared understanding of cultural heritage information by providing a common and extensible semantic framework that any cultural heritage information can be mapped to."

Currently, and in the framework of a Call for Tenders, the Greek Ministry of Culture has launched the implementation of a special mediator service through which the Greek Cultural Assets which have been digitised and documented so far are semantically unified to enable single point access to the wealth of these vast information sources.

Homer's Illiad, VIII, 245-253, codex F205, late 5ct or early 6ct

Users' view

HINA – digitising and exploiting the news archives

Željko Hodonj
Development consultant in HINA

In 2008 the Croatian News Agency (HINA) became the focal point for the development of digital archives in Croatia. Only a year earlier the Croatian parliament entrusted HINA with the protection of (news media) archives containing some 14 million documents originating from two databases, the archives of newspapers articles in the Vjesnik publishing house (VND records) and the EVA database, created by HINA. These databases have the status of protected archive collections of Croatian culture.

Being a news agency, HINA's prime mission is to provide credible and unbiased information about events and people as soon as possible and deliver it to its subscribers. As a part of its business, HINA has planned the development of a multimedia database aimed to create new services for its subscribers. It has prearranged its workflow for creating new business packages through media monitoring services. The project now includes all relevant printed news media sources in Croatia and all news production created by HINA. Simultaneously, we are going to begin with digitization of the VND records, working backwards from the most recent ones.

This is a simple, yet a demanding approach, to link all articles from recent printed media and articles created in the past, because it is the only model that would enable a true understanding of the content and detecting the trends; a dynamic connection between time and topics, individuals and events.

In this context a new general objective has also been defined; to support information access to all Croatian citizens and enable them access to media information at the lowest cost possible.

The digital archive as an area of organized knowledge has been designed as a system that connects information and events, people and their activities; a system which supports the organization of knowledge concerning the connection between people and events, people and people, event and event – as a system which recognizes and builds up a network of relations and correlations between topics and the time of their occurrence. The preconditions for achieving this objective and all of its segments was to find a partner capable of supporting the project with expert knowledge.

In April 2008, HINA opened a public tender to select the best provider for the services of semantic analysis of the text in a modular fashion. A joint offer submitted by Faculty of Electric Engineering and Computing and the Faculty of Humanities and Social Sciences in Zagreb was selected in this tender as the best provider. By late 2008, the laboratories of the two faculties will provide prototypes: a module for lemmatization and morphosyntactic tagging of Croatian words, a module for recognizing and classifying named entities in the Croatian language, a module for automatic document classification, a module for automatic detection of key words from documents in Croatian, a shell that enables access to each module via a web service and a system for monitoring the work of all modules.

By the end of April 2009, all versions of the system will be tested thus bringing HINA to the first level of the realization of its objectives.

This approach illustrates how HINA understands relations between pragmatic necessities of its business interests and the protection cultural heritage and legacy knowledge, how it contributes to the protection of cultural entities and European values.

Editors' note

On this page we publish opinions, discussions, view and arguments that will usually come from two sides. One will illustrate the standpoint of CLARIN users or "consumers", while the other will try to present the ideas that are coming from the direction of LRT developers.

Developers' view

Which LR&T are available for supporting HINA?

Bojana Dalbelo Bašić
University of Zagreb, Faculty of Electrical Engineering and Computing

The tender for providing the services of semantic analysis of the Croatian newspaper texts for Croatian News Agency (HINA) had put a serious challenge before us. Since the tasks were many and covered different areas of expertise (language resources and technologies, together with knowledge technologies and their computational realisation), Faculty of Humanities and Social Sciences (Department of Linguistics) and Faculty of Electrical Engineering and Computing (Department of Electronics, Microelectronics, Computer and Intelligent Systems) submitted a joint offer to HINA. The architecture of the system that we had to provide solutions for was modular so different modules should be used for different simple tasks and that more complicated task could be achieved by their recombination or usage in different workflows. The tasks that we have to deal with are ranging from the simplest to the more complicated ones.

The basis of the system represent the LR&T tasks such as lemmatization and POS/MSD tagging of Croatian texts which has only been done so far in academic and experimental environments and was never tested in real, industrial strenght conditions where it has to achieve the highest possible rate of accuracy without sacrificing the speed, which can be crucial in newswire text flows. For this purpose a new hybrid tagger for Croatian was developed which reached around 97% accuracy. The module for lemmatization and full morphosyntactic tagging will also be used later for query processing during the user access to textual databases.

The next task, which builds on the previous one, is named entity recognition and classification. The system for NERC in Croatian texts was developed three years ago (with an f-measure of around 90%) but also as a research prototype with only few tests in real business applications. At this first stage only names of persons, organizations and locations will be recognized and classified according to the existing gazetteers, while later this initial classification will be used for more elaborated schemata.

The module for document classification has to deal with the daily flow of news documents that measures in tens of thousands. It will be able to classify new documents to a predefined classifying scheme. The initial tests without the thorough training of the classifiers yielded 86% f-measure which was considered excellent.

The module for keyword detection is yet to be defined in detail but the overall idea is to submit each document to statistical processing that will measure statistics of individual single- and/or multiword units within the parts of the document to the whole document or to the whole document collection. The significance of keywords will be measured statistically and positionally (i.e. within the document).

The module for monitoring should provide opportunity to adapt the functioning of other modules (i.e. adding new named entities in gazetteers, retraining the classifiers, error detection/correction etc.). This basic system will allow further development of more complicated knowledge-oriented services such as social networking. C

The ISO standards for LR&T that are being developed

Thierry Declerck
DFKI GmbH

The production, processing, use and re-use of multilingual linguistic data constitute a time-consuming and costly part of the daily work in the language industry. There is a critical need for established standards to enable the interoperability and re-use of multilingual data and linguistic processing tools in order to facilitate processes such as harmonisation, localisation, machine translation and cross-lingual information retrieval.

The need to provide industry-validated standards for language resource management has been recognized within the ISO (International Organization for Standardization) community for a long time. A special ISO Technical Committee was established more than 50 years ago, dedicated mostly to terminology and terminological resources.

In a review of the first 50 years work of this committee, it is declared that: "…terminology plays a crucial role wherever and whenever specialized information and knowledge is being prepared (e.g. in research and development), used (e.g. in specialized texts), recorded and processed (e.g. in data banks), passed on (via training and teaching), implemented (e.g. in technology and knowledge transfer), or translated and interpreted. In the age of globalization the need for methodology standards concerning multilingual digital content is increasing ".[1]

Recently, in 2001, the name of the Committee has been changed into "Terminology and other language resources", and this step was crucial for many partners involved in CLARIN since it opened the standardisation work to all academic and industrial activities dealing with language in one form or another.

So, for example, the development of linguistically annotated corpora or computational lexicons can now be accompanied by ISO standardization efforts.

Four Subcommittees of TC37 have been established, the latest one being TC 37/ SC4 "Language resource management", created in 2002 in cooperation with ELRA, the European Language Resource Association, which is also a partner of CLARIN, and a main organiser of the LREC conference. The general overview of TC 37 is now:

– ISO/TC 37/SC 1 "Principles and methods"
– ISO/TC 37/SC 2 "Terminography and lexicography"
– ISO/TC 37/SC 3 "Computer applications for terminology".
– ISO/TC 37/SC 4 "Language resource management"

Visit the home page of TC37/SC4 at http://www.tc37sc4.org
for more information on LR&T ISO standards that are being developed

There are six development levels for ISO standards, described below, based on the definitions in the glossary of the ProSTEP iViP Association:[2]

1. NWI: New Work Item. This document contains the proposed scope and contents of a newly suggested topic for standardization. If there is sufficient interest (of at least five nations i.e. national representatives) in this subject, the document will be advanced to a working draft.

2. WD: Working Draft. The Working Draft document is derived from the NWI and contains the essential technical contents of the future standard. This includes the coordination of different technical approaches.

3. CD: Committee Draft. This document is the first version of a part of the future standard, which is internationally voted.

4. DIS: Draft International Standard. This level emerges from the CD stage through agreed changes and additions.

5. FDIS: Final Draft International Standard. The Final Draft International Standard is an optional stage in the standardization process. Normally, the FDIS stage is only reached if changes to the DIS document before publication as IS are necessary.

6. IS: International Standard. After being approved, a (F)DIS finally becomes an International Standard under ISO copyright. It thus becomes an accepted document, which provides a reliable basis for implementations and harmonised solutions.

Currently, the standards being discussed in TC 37/SC 4 concern the markup of lexical resources and the annotation of linguistic data at various language processing levels (morpho-syntax, syntax and semantics). For this purpose a general framework (Linguistic Annotation Framework, LAF) defining the representation of these phenomena is in development, with the purpose to enable interoperability between the various types of markup and annotation information involved.[2]

Many partners of CLARIN can and will contribute to those standards, in areas such as the citation of resources, interoperability of language data and tools.

But for CLARIN it is necessary that the LR part of the whole infrastructure stick as much as possible to the existing standards. Otherwise it would be much harder to achieve our general goal.

[1] See http://www.infoterm.info/pdf/activities/Standing_document_02_50_years_ISO_TC_37.pdf

[2] See http://www. prostep.org/nc/en/metanav/glossary.html

[3] See also http://www.tc37sc4.org/ for more information.

CLARIN at LREC 2008

Steven Krauwer
CLARIN coordinator

For those of you who attended LREC 2008 in Marrakech it was hard to miss CLARIN. No less than 12 papers at the main conference contained references to CLARIN, even if the project itself had only recently started, well after the submission deadline for papers for the main conference. For papers presenting results emerging from the CLARIN project we will of course have to wait until the next LREC conference in 2010.

In Marrakech the focus was on our plans (and concerns) for the future, on mobilizing the community and on identifying opportunities and parties in humanities and social sciences for future collaboration.

Apart from a general presentation of the CLARIN project at the main conference, which shouldn't contain too much new information for most of our Newsletter readers, we had a special panel session dedicated to key issues in building an infrastructure for language resources and tools. These issues included the following

1) What is the scope of building a common language resources and tools infrastructure and what is the particular role that endangered and minority languages should play in this process?
2) How can language resources and tools best be made available and made accessible to the research communities at large?

For example, resource providers may be fearful of what happens to "their" resources and tools. Will they be asked to hand them over to some big "data bureaucracy"? How can accessibility and IPO issues be reconciled? What technical solutions are available for all of this? In this panel session, moderated by Erhard Hinrichs (University of Tübingen), five panellists highlighted different aspects. Sadaoki Furui (Tokyo Institute of Technology) gave an overview of what was happening in Japan, especially in the program "Framework for Systematization and Application of Large-scale Knowledge resources". Sebastian Drude (Museu Goeldi Belem / Freie Universität Berlin) focused on digital language archives for minority and endangered languages and gave an overview of the major language documentation programs and the special characteristics of archives for endangered languages. Nicoletta Calzolari (CNR-ILC Pisa) looked at the language resources landscape from the FlaReNet perspective, a new EC project, very closely related to CLARIN, that aims at providing recommendations for future language resources policies in Europe (see http://www.ilc.cnr.it/viewpage.php/sez=ricerca/id=904/vers=ita). The interesting difference between CLARIN and FLaReNet is that where CLARIN's target audience is the social sciences and humanities research community at large, FLaReNet covers the whole spectrum of parties interested in language resources and technology, ranging from basic research to industrial development. Peter Wittenburg (Max Planck Institute Nijmegen) discussed the typical pillars of the technical CLARIN infrastructure and in my own presentation I identified a number of possible ways for CLARIN to fail. I don't think that we managed to solve any of the major problems during the discussion that followed, but it was interesting to see that there was a keen interest from the audience in CLARIN and in infrastructure issues in general. The slides of the panel presentations and of the presentation at the main conference are available on the CLARIN website at http://www.clarin. eu.

On the whole I must say that many of the sessions and discussions in the main LREC programme and in the workshop programme showed that infrastructures are a hot topic, not just in Europe but all over the world. This calls for intensive international collaboration in order to make sure that we all move in the same direction and, more importantly, that we move towards common standards to ensure maximal interoperability. During the conference we also managed to organise an internal meeting for CLARIN consortium partners attending the conference. As our project budget for travel is limited we will try to organise CLARIN meetings in the margin of some of our major conferences and we hope that at the next LREC conference there will be room for us to organise another meeting, but then not just for consortium partners but for everybody actively involved in CLARIN, both in the EC project and in the accompanying national projects launched to prepare the construction of the CLARIN infrastructure at the national level. C

The CLARIN LREC2008 meeting

Digital Humanities 2008

Oulu, Finland, June 25-28th

Martin Wynne
CLARIN EB member

Digital Humanities is the joint international conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, and has been taking place annually since 1989, since 2006 under the banner of 'Digital Humanities'[1]. It claims to be the oldest established meeting of scholars working at the intersection of advanced information technologies and the humanities. This year the conference was held at the University of Oulu in Finland, and Tamás Varadi and Martin Wynne presented a poster outlining what CLARIN was all about.

Many presentations at DH2008 were of relevance to CLARIN. These included CenterNet,[2] an initiative to build a network of digital humanities centres, originating in North America, but now with participation from around the globe. A panel session on the opening morning discussed aspects of 'Defining an international humanities portal' in the context of CenterNet. It soon became clear however that what was needed was not another portal to compete for space on our desktops with those of our faculty, institution, national community, academic discipines, etc. What we need is to present our resources and tools as standards-conformant services so that they can be integrated into the researcher's environment.

Also emanating from this community, and worthy of attention from CLARIN members, are TaporWare, a set of online text analysis tools[3], Project Bamboo, an initiative to develop new shared technology services for the Humanities[4] and Heurist, a free online database service developed for scholars in the Humanities[5]. In fact it has become clear that the priority is not for more tools, more portals, or more initiatives, but a way of connecting together the existing fragmented landscape into a ecosystem where creators of resources, tools and services can do so in such a way that they can interoperate with the other resources relevant to their users. The CLARIN approach is to promote this interoperability and to faciliate their deployment within a sustainable infrastructure.

DH2008 is effective as a gathering of a group of scholars who are profitably sharing ideas, techniques, resources and tools across traditional discipline boundaries. It was a subject of ongoing debate among scholars at this conference whether they represent more than this, and whether they should see 'humanities computing' (or nowadays more often 'digital humanities') as an academic discipline in its own right. Whatever the answers to this question it is clear that it will be very useful for CLARIN to engage with the discussion.

In the final plenary session, invited speaker Sylvia Adamson of the University of Sheffield demonstrated how the historical linguistic research in which she is engaged would benefit from better electronic resources and tools, and posed a challenge to "the geeks in the audience" to develop a more effective working environment for scholars like her. This is precisely the challenge that CLARIN has accepted.

[1] DH2008: http://www.ekl.oulu.fi/dh2008/

[2] CenterNet: http://www.digitalhumanities.org/centernet/

[3] TaporWare: http://taporware.mcmaster.ca/

[4] Project Bamboo: http://projectbamboo.org/

[5] Heurist: http://heuristscholar.org/

ESF meeting on humanities in CEE countries

Marko Tadić
Editor

The special European Science Foundation topic-oriented worksop organized by the Standing Committee for the Humanities (SCH) was held between 26th and 28th June 2008 in Sofia.

Since the exact title of the meeting was Stakeholder Workshop "Central and Eastern European Scholarship in the Humanities – harnessing the assets", one could ask him/herself was it relevant for CLARIN in any respect. It certainly was since the CLARIN research infrastructure FP7 project was presented in an invited talk by Marko Tadić.

The meeting assembled some 50 invited participants from 12 countries and ESF officials – representing some of the most relevant institutions in Central and Eastern European countries among the ESF membership EE, LT, LV, PL, CS, SK, HU, HR, SI, RO, BG and important partners elsewhere – to debate the necessary measures to further strengthen the participation of CEE scholarship in the European Research Area. The workshop theme and philosophy was "Harnessing the Assets"; it was understood that efforts need to be made to better promote the strengths of CEE Humanities scholars as research partners. While different presentations also highlighted structural obstacles that still exist in this process, it was clearly demonstrated that the rhetoric of "catching up" has to be abandoned since there is nothing to catch up between CEE countries and Western Europe countries. The CEE countries are simply in some respect different and this difference has to have its reflections in Humanities (and Social Sciences) research. The historical starting points are not the same, the financial support is not the same, even the position of Humanities research is not the same in CEE and Western Europe countries. This differences were discussed. The workshop also addressed the issues of international comparability of Humanities scholarship in smaller linguistic communities, and the related challenges of internationalization and globalization. Nevertheless, the workshop was looking for an agreement on a joint strategy to move forward, comprising as possible elements of study:

- commissioned surveys (by country and/or by discipline);
- identification of pilot initiatives;
- involvement of Institutes of Advanced Studies and university-based centers of excellence;
- dialogue with neighboring sciences (e.g.: social sciences; environmental sciences);
- consortium building (third-party-funding);
- role of ESF member organizations in promoting CEE academic assets.

The role of a research infrastructure for Humanities was strongly put forward and in this respect CLARIN gained the deserved attention not just by its presentation but also in discussions during the whole duration of the workshop. The idea of LR&T as supporting technology for text(-based) sciences that would allow the development of new Humanities paradigm (e-Humanities) was appealing to the majority of participants while some of them still showed some reserve to its final shape. The scepticism was not only based on the problems of technical nature (i.e. organizing the federation of archives, succesful application of grid technology, open and permanent access to research data and results etc.) but was also based on the problem of empirical vs. rationalistic approach in Humanities research in general (e.g. Chomsky's philosophical standpoints) which is a well known dichotomy that certainly will not be solved soon.

CLARIN's role in open and permanent access to research results and data

Peter Wittenburg & Tamás Váradi
CLARIN EB members

The topic that was introduced in the previous issue of the CLARIN Newsletter (CNL) raised quite a few questions. Therefore CNL has organized an interview with Peter Wittenburg and Tamás Váradi in order to clarify and elaborate further their standpoint on the role of LR&T and its availability for researchers and society in general.

CNL: In your article in the previous CNL issue you sketched that there will be a competition between some big companies, currently abbreviated with the buzzword Amazoogle, on the one hand and infrastructure initiatives, such as CLARIN, on the other hand. Why are you sceptical about their services?

W&V: Our skeptical attitude emerged from the experiences with the big publishers who broke the "old" deal with the research communities and began to see scientific content primarily as a market and not as a service. Google's European vice-president stated again very clearly that their primary concern is business, of course, and we should not be naive in this respect. Prices are primarily not dependent on the production costs, but on the market opportunities, i.e. the research community needs to be careful not to find themselves again in a situation where they need to pay too high prices to access the content which they produce themselves. Too strong dependencies without alternatives don't seem to work.

CNL: Isn't there also a chance of collaboration between complementary services?

W&V: Of course, we could imagine a complementary situation between the Amazoogles and, for example, CLARIN and we already tried to contact the big companies to establish collaborations. This complementary nature has been pointed out by the British JISC (Joint Information Systems Committee) initiative already. However, everything will very much depend on how the big companies will position themselves and on the policies they will follow: whether they decide to pursue a more collaborative or a more ignorant policy. If CLARIN, for example, would be allowed to link up to the emerging huge digital resource base of the companies, then this would be beneficial for the user community and certainly we would not hesitate to allow them access to our resources. What would be important is that there are proper agreements that are beneficial for all sides.

CNL: Aren't the Amazoogle services the best answer to overcome the huge fragmentation we are suffering from in the humanities? After all, as history has shown, they could ultimately enforce a standardization with respect to the complex rights and formats issues due their monopoly role. Wasn't it Microsoft who pushed us forward to a quasi standard at operating system level?

W&V: It could indeed happen that CLARIN does not have sufficient power in our community to convince researchers to make use of certain standards and to simplify the rights situation. In that case the Amazoogles could succeed. They could simply rely on the fact that everyone who wants to be seen in the web needs to adhere to the constraints or practices imposed by them. The pressure could become high for each researcher so that all concerns are set aside. Basically this implies that CLARIN also has to push its own community in the direction of harmonization and simplification. On the other hand we need to remain sensitive enough to cope with special requirements.

CNL: So no reason for CLARIN, DARIAH, ERIH, etc. to come up with an alternative model?

W&V: For sure we need to learn from the Amazoogles in many ways. Again we count on a collaborative attitude. One of the aspects where we need to learn is how to efficiently establish and run a network of LRT centres. For several reasons, centralization could be very beneficial – think only of the potential saving in energy costs to run big servers. But there are so many reasons that require a more distributed approach. Think of the relevance of "national" centres as being responsible for "their" languages. We know how important identification is for the amount of effort and money people are willing to spend.

CNL: Do we understand correctly that you don't see any technical motivations for a distributed CLARIN future?

W&V: No, of course there are also technical reasons: let us mention for example the need for distributed storage to ensure data survival and distributed services to guarantee high availability and performance. Let us point out that the Amazoogles also need a distributed approach to storage and services for similar reasons. However, that does not commit them to a model of distributed decision making. There is still a single decision taking board at the top of the hierarchy and this board can decide about optimal solutions to make good business. The situation is different with CLARIN from this respect. We are committed to other aspects than "business success" such as geographic distribution and balance of other kinds, which also means that no one should expect that we can be as cost efficient as the Amazoogles are trying to be.

CNL: When you say that you could imagine an atmosphere of collaboration, do you have an example in mind?

W&V: Yes, for sure. The publishers, for example, have taken over the job to produce and distribute books representing much of our cultural heritage. But, in addition, each country has a national archive and a national library. Their role is beyond the one of publishers. They need to take care of completeness, archiving, curation, etc. and for these aspects there is no self-supporting business model, but a national interest. With respect to the digitized heritage, including language resources, this necessity of splitting roles is even more important. Still we have a highly dynamic situation where the various roles and business models are not yet settled but are developing and influencing each other and also shaping the landscape.

CNL: Could you imagine already now an aspect of work where you think that the roles will be different and therefore complementary?

W&V: The Amzoogles are companies that come and go, in particular, nowadays, while the lifetime of states or cultural regions in general seems to be much longer. States and regions have an interest in taking care of long-term preservation of their heritage and culture. So, with respect to language resources, it will be the task of the linguistic community, in collaboration with the states or regions, to take care of their long-term

preservation. The Amazoogles can't give guarantees beyond their business models.

CNL: Are there other aspects where you see differences?

W&V: Good and efficient business will have problems in taking care of all sort of special needs such as minority languages, special formats, special rights, etc. These areas where special knowledge is required, where the number of users will be limited – usually specialists – and where special concerns and regulations will increase the costs. Take as an example the amount of curation that is necessary to bring existing lexicons into a generic format that will allow even laymen to determine cognate sets, etc. It requires deep knowledge about the language, the encodings used and the structural characteristics to carry out useful transformations and this would only be useful for a small community of potential users in education and research. We don't see a business model for this.

CNL: In all your argumentation you are indicating that there could be much gain in collaboration. Why did you put the "competition" aspect in the foreground of your argumentation in your statements?

W&V: Indeed a good question. Of course, on our side the competition is a challenge to work at efficient cost levels and to reduce complexities where possible. The other part has to deal with sensitivity of various sorts. Funding agencies may easily come to the conclusion that the Amazoogles will take care of all aspects of language resources and technology, so that they don't have to care. We wanted to stress that this is not true. In contrast, there is a danger that funders may want to reduce funds due to the "cheap" offers of the Amazoogles. Many researchers are charmed by the simplicity of the Amazoogle user interfaces. However, for deep linguistic or humanities research the options are not sufficient. And, as we indicated, the financial questions remain.

CNL: OK let's discuss the correctness of your expectations in this respect. Offering reliable services will cost some money whoever is giving them. Why should a company not ask for some usage or access fee?

W&V: That's absolutely correct. Whoever will give services, they will cost some money. It is completely ok that a company asks a certain price for a certain service. Currently the Amazoogle offer to store data, for example, free of charge, they currently only house data that is open access and they have their license terms, so depositors have to adhere to these very simple rules. But, come on, it is business, i.e. when there is an option to make money, a company would do so. So even if storage is free, it may and will be the case that access in whatever form will be associated with fees and as we said earlier: monopolies tend to base their prices not on the actual costs but on what the market can give. And the Amazoogles are monopolies that could define prices.

CNL: So you think that the cost effectiveness of their services will not lead to low service prices?

W&V: Correct – the cost effectiveness will only be offered as low prices to the customers if there is real competition. So, here is an argument for having initiatives such as CLARIN to create a competitive domain.

CNL: Some say that if the Amazoogles would change their cost model, immediately commercial competitors would show up – so they cannot change their prices if they do not want to fail.

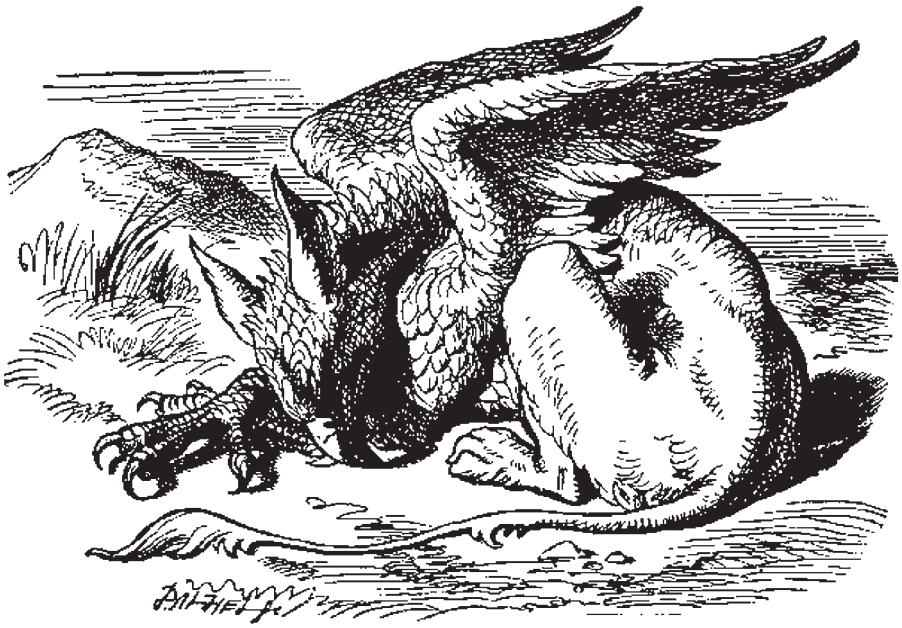
W&V: As long as there is competition, as you stated it. So research infrastructure initiatives could play a role in this game. But the web opens up a new phase in centralization, as the Amazoogles are demonstrating. They offer excellent services and the whole world is adapting to their style and/or content of services. Other service providers are reduced to marginalities or, if you want, only a few will remain. If the major market players agree on terms, we will be dependent on their services and it will become almost next to impossible for newcomers to enter the market.

CNL: It is a known fact that the Amazoogles are making money mainly exploiting advertisements – so this is another reason not to change their business model.

W&V: Well, we already discussed this aspect. After all, it is business. Who can predict the precise circumstances for making money in 5 years?

CNL: So, you remain sceptical.

W&V: Yes. We are all using Amazoogle services and we like them. But let's be careful with the scientific content and not make research data and research results dependent on the business models of big monopolies. Let the research communities determine how they want to access and use scientific data.



Could such impossible combinations like Amazon and Google taken together live at all? Gryphon illustration by Sir John Tenniel for Lewis Carroll's *Alice in Wonderland*

CNL: What if the Amazoogles established scientific advisory boards where researchers would be invited to determine the rules of the game at least for scientific usage?

W&V: If it were to be guaranteed that even their Executive Boards would accept the advice of the researchers, it would be a big step but unfortunately this is not the way how business is functioning.

CNL: A final statement please.

W&V: Everybody agrees now that it is good to have the Amazoogles as they keep driving us ahead in many ways. But we need research infrastructures such as CLARIN as well to support research, to take care of specialties and establish at least some form of competition.

There is also much space for collaboration, as well. **C**

The Danish CLARIN project

Hanne Fersøe
Centre for Language Technology, University of Copenhagen

Along with other governments, the Danish government has decided to invest in research infrastructures for all fields, including the humanities and social sciences as a part of the Danish Globalization Strategy. The infrastructure initiative has a total budget of 600 million DKK (appr. 80 million Euro) which will be distributed with 200 million per year over 3 years, 2008-2010. Funding is not limited to projects which are included in the ESFRI Roadmap, so infrastructure projects of all kinds may apply for funding.

A consortium headed by the University of Copenhagen has been given a three year grant of 15 million DKK (appr. 2 million Euro) to construct a Danish research infrastructure for the humanities integrating written, spoken, and visual records into a coherent and systematic digital repository. The project runs from January 2008 until the end of 2010. The budget was granted following a call for expressions of interest, where our consortium was invited to submit a full proposal. Following an international evaluation with three reviewers, the proposal was selected for funding with a 50% budget cut.

The Danish CLARIN project was accepted as the only proposal solely for the humanities, and the factors that supported it were probably numerous: the consortium is strong and to the point with four universities and four cultural institutions, thus providing a very good basis for a true humanities project, and at the same time with the technical skills necessary. Furthermore, there is an understanding in the Danish research administration that the humanities need infrastructure support, in particular IT support; the Ministry conducted a hearing about the ESFRI projects among all the relevant and interested parties in Denmark, and it is assumed that CLARIN must have received good support; the Centre for Language Technology is a well-known player in the field; the technical as well as the management part of the proposal were well accepted by the reviewers; the EU CLARIN project had been invited to contract negotiations by the time the full proposal was submitted.

The partners include eight leading Danish humanities institutions: four universities, a university library, a museum, and two government research institutions. The ten consortium members are:

University of Copenhagen with three departments from the Faculty of Humanities:

1) Centre for Language Technology – coordinator of the project,
2) Danish National Research Foundation Centre for Language Change in Real Time,
3) Department of Scandinavian Studies and Linguistics,
4) University of Southern Denmark,
5) University of Aarhus,
6) Copenhagen Business School,
7) The Royal Library,
8) The National Museum of Denmark,
9) Society for Danish Language and Literature,
10) Danish Language Council.

Mission and Vision

The vision is to create a researcher's toolbox by establishing a number of digital Danish text, speech and visual resources and associated tools and to integrate these resources into a web-based environment for research thus creating a much needed support for Danish humanities and enhance its possibilities for European collaboration. The Danish CLARIN project will also improve the conditions for Danish language technology research and development by starting a structured approach to a Danish BLARK.

The Danish CLARIN project will follow standards and recommendations developed in the preparatory phase of the European CLARIN project, but it is important to realize that this project has not been granted as a preparatory phase project in parallel with the European project. It involves an independent Danish investment in the construction of a national infrastructure that will stand alone as a vital contribution to the Danish research enterprise. For this reason it was vitally important for the consortium to design and plan the activities in such a way as to be able to deliver not only the technical infrastructure as a result, but also as many types of content as possible.

Activities

The work is organized in thematically defined groups of activities, three of which are dedicated to making content available, while one focuses on the technical infrastructure.

The content that will be made available to the research communities comprises both existing basic written language resources (contemporary and historic, general language and specialised sublanguages, literary and professional, as well as parallel corpora with Danish as one of the languages), three different spoken language corpora and associated tools, and data resources such as traditional and electronic dictionaries, dictionaries and semantic word nets meant for computer systems, and the linking between different dictionaries as well as between dictionaries and corpora.

The technical infrastructure will include a single web user interface to serve as the Danish CLARIN platform. This platform will give access to all the tools and text resources of the infrastructure, as well as a personal workspace, communication facilities, user authentication and rights management, and search and retrieval facilities. C

D-SPIN – the German CLARIN Initiative

Erhard Hinrichs, Peter Wittenburg, Alexander Geyken, Lothar Lemnitzer, Andreas Witt

As a support action for CLARIN the German Federal Ministry of Research and Education (BMBF) has launched D-SPIN ("Deutsche Sprachressourcen-Infrastruktur" – Infrastructure for German Language Resources), an infrastructure project of language resources and tools. Apart from D-SPIN, the BMBF is also funding two additional eHumanities projects: Textgrid and eAQUA.

All these projects will closely collaborate and will coordinate the necessary steps for the construction of an eScience infrastructure for the humanities. On the international level, D-SPIN will collaborate with DOBES ("Documentation of Endangered Languages", http://www.mpi.nl/DOBES) and BABEL ("Better Analysis Based on Endangered Languages", http://www.esf.org/activities/eurocores/programmes/eurobabel.html).

The D-SPIN project is co-ordinated by the University of Tübingen. The partners are: the Berlin-Brandenburgische Akademie der Wissenschaften in Berlin, the DFKI in Saarbrücken, the Institut für Deutsche Sprache in Mannheim, the MPI for Psycholinguistics in Nijmegen and the universities of Frankfurt, Gießen, Leipzig, Stuttgart and Tübingen. The project will invite all language resource providing institutions in Germany to become involved in the project at an early stage. To this end D-SPIN will organize a national workshop with all interested institutions in the spring of 2009.

D-SPIN will run concurrently with CLARIN and will mirror its structure in order to facilitate the cooperation between both projects and in order to generate synergies. However, most of the work is devoted to national tasks, in particular to the establishment of service centers and the localisation, description and integration of key language resources and tools in Germany.

The user communities targeted by D-SPIN are not only limited to linguists and language technologists, but also include other humanities disciplines. Since it is not possible to address all relevant subfields of the humanities at this stage, D-SPIN will try to make tools and resources available that are of interest to scholars from the field of history.

In order to identify user needs, D-SPIN will organize a workshop with board members of the largest on-line community of historians ('H-Soz-Kult'). Since libraries are increasingly playing the role of historical resource providers, D-SPIN will also seek the cooperation of the Staatsbibliothek zu Berlin (Berlin State Library). The "Sammlung Erster Weltkrieg" (text archive of World War I), a project of the Berlin State Library, will serve as a test case for investigating the usefulness of different levels of linguistic annotation, e.g. lemmatisation and named entity annotation. Additional possibilities include annotations by the scholars themselves in the form of stand-off annotation.

The exploration of such usage scenarios for the field of history will help to identify specific requirements for the construction of an eScience infrastructure which will ideally can be applied to other humanities disciplines as well. For further information about D-SPIN, please consult: http://www.sfs.uni-tuebingen. de/dspin.

CLARIN-DRIVER Proposed Collaboration

Dale Peters (DRIVER), Peter Wittenburg (CLARIN)

Recently at the DRIVER Summit in January 2008 it became obvious that DRIVER and CLARIN are complementary infrastructure initiatives that will benefit from collaboration and take profit from each others work. While CLARIN is a research infrastructure bringing together language resources and technology, DRIVER is oriented at a broader scope to harvest data from across disciplines. Here we want to briefly present the DRIVER project and discuss possible collaboration elements.

DRIVER

The primary objective of DRIVER is to establish a flexible, robust, and scalable infrastructure for all European and world-wide digital repositories, managing scientific information in an Open Access model, increasingly demanded by researchers, funding organisations and other stakeholders. DRIVER's mission is to expand its content base with high quality research output, including textual research papers and other scholarly publications. The recent D-NET v1.0 open source software release forms the basis of this distributed service-oriented architecture that enables enhanced interoperability of data and service-providers, with functionality ranging from search, recommendation, collections, profiling to innovative tools for repository managers. By building a robust network of content providers, enhanced with the set of services and software tools that DRIVER offers, the DRIVER infrastructure also enables further collaboration to domain specific research communities in a co-ordinated network of repositories from a growing number of institutional repositories and from national institutions and data aggregators in Europe. Future developments will see the extension of the infrastructure to global research communities in a vigorous awareness and advocacy programme fostering the development of digital repositories. The DRIVER network offers a support service for repository managers, a dynamic set of guidelines aimed at data interoperability, and the strategic support for new forms of scholarly communication.

Collaboration

CLARIN centres and aggregated services will form nodes in the robust network of DRIVER content providers. It will offer its metadata descriptions so that DRIVER can harvest them and offer them in its services. Also services of DRIVER repositories should become nodes in the CLARIN network of language resource providers to make the data available for researchers.

The proposed collaboration between CLARIN and DRIVER promises a joint European research infrastructure for comprehensive service to the humanities disciplines with respect to language resources and technology and therefore plays a key role in the construction of an efficient research and innovation environment.

CLARIN calendar of events

Here is the list of CLARIN events and events from the fields of language resources and language tools that could be of an interest to CLARIN members.

August 2008

2008-08-04 to 2008-08-15: ESSLI 2008, Hamburg, Germany

2008-08-16 to 2008-08-24: COLING conference with pre- and post-conference tutorials and workshops, Manchester, UK

September 2008

2008-09-08 to 2008-09-12: TSD2008 conference, Brno, Czech Republic

2008-09-11 to 2008-09-12: FSMNLP2008, JRC Ispra, Italy

2008-09-17 to 2008-09-19: CLEF2008, Aarhus, Denmark

October 2008

2008-10-14 to 2008-10-15: eScience Workshop: Metadata Principles, Berlin, Germany

November 2008

2008-11-01 to 2008-11-03: Chicago Digital Humanities/Computer Science Colloquium, Chicago, USA

2008-11-10 to 2008-11-11: Web services architecture in CLARIN, Munich, Germany

December 2008

2008-12-09 to 2008-12-10: European Conference on Research Infrastructures, Versailles, France

2008-12-17 to 2008-12-12: IEEE e-Humanities Workshop, Indianapolis, USA C

Join CLARIN

CLARIN project is a combination of Collaborative Projects and Coordination and Support Actions, registered at the EU under the number FRA-2007-2.2.1.2. It started with the preparatory phase in 2008 that will make the grounds for the next phases and it will cover the generic, language independent activities. In order to do our work properly we will have to rely on a much wider circle than just the formal consortium partners in the project. For this reason we have opened up all our project working groups for participation by organizations that are not part of the consortium.

Members

Country; Institution; Location; Contact person

Austria: University of Vienna; Vienna; Gerhard Budin

Belgium: ALT (Acquiring Language through technology); Leuven – Kortrijk; Hans Paulussen

Center for Computational Linguistics ; Leuven; Ineke Schuurman

Center for Dutch Language and Speech, University of Antwerp; Antwerp; Walter Daelemans

ELIS-DSSP; Gent; Jean-Pierre Martens

Legal Informatics and Information Retrieval, Katholieke Universiteit Leuven; Leuven; Marie-Francine Moens

Laboratory for Digital Speech and Audio Processing – VUB – ETRO/DSSP ; Brussels; Werner Verhelst

ESAT-PSI/Speech; Leuven; Patrick Wambacq

Bulgaria: Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences; Sofia; Svetla Koeva

Institute for Parallel Processing; Sofia; Kiril Simov

Mathematical Linguistics Departement, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences; Sofia; Ludmila Dimitrova

Croatia: University of Zagreb, Faculty of Humanities and Social Sciences; Zagreb; Marko Tadić

Institute of Croatian Language and Linguistics; Zagreb; Damir Ćavar

Cyprus: Cyprus College / Research Center; Nicosia; Antonis Theocharous

Czech Republic: Charles University; Prague; Eva Hajičová

Faculty of Informatics, Masaryk University ; Brno; Aleš Horák

The Institute of the Czech Language, Czech Academy of Sciences; Prague; Karel Oliva

Denmark: Center for Sprogteknologi, University of Copenhagen; Copenhagen; Bente Maegaard

Dansk Sprognævn – Danish Language Council; Copenhagen; Sabine Kirchmeier-Andersen

Society for Danish Language and Literature; Copenhagen; Jørg Asmussen

Estonia: University of Tartu; Tartu; Tiit Roosmaa

Finland: CSC – the Finnish IT Center for Science ; Espoo; Tero Aalto

University of Helsinki; Helsinki; Kimmo Koskenniemi

Department of Foreign Languages and Translation Studies, University of Joensuu; Joensuu; Jussi Niemi

University of Tampere; Tampere; Eero Sormunen

The Research Institute for the Languages of Finland; Helsinki; Toni Suutari

France: ALTIF; Nancy; Bertrand Gaiffe

TELMA/DIS CNRS; Paris; Florence Clavaud

CNTRL; Nancy; Bertrand Gaiffe

Evaluations and Language resources Distribution Agency (ELDA); Paris; Khalid Choukri

Université Paris 4 Sorbonne / CELTA ; Paris; Andre Wlodarczyk

LIF-CNRS ; Marseille; Michael Zock

Germany: Berlin-Brandenburg Academy of Sciences; Berlin; Alexander Geyken

Deutsches Forschungszentrum für Künstliche Intelligentz; Saarbrücken; Thierry Declerck

Institut für Deutsche Sprache; Mannheim; Marc Kupietz

Max Planck Institute for Evolutionary Anthropology; Leipzig; Hans-Joerg Bibiko

University of Frankfurt/Main Comparative Linguistics; Frankfurt/Main; Jost Gippert

University of Leipzig; Leipzig; Codrina Lauth

University of Stuttgart; Stuttgart; Ulrich Heid

Universität Tübingen; Tübingen; Erhard Hinrichs

University of Giessen; Giessen; Henning Lobin

Computational Linguistics Department, University of Heidelberg; Heidelberg; Anette Frank

University of Augsburg ; Augsburg; Ulrike Gut

Greece: Institute for Language and Speech Processing; Athens; Stelios Piperidis

Hungary: Academy of Sciences; Budapest; Tamás Váradi

Budapest University of Technology and Economics Media Research (BME MOKK); Budapest; Peter Halacsy

University of Szeged, Department of Informatics, Human Language Technology Group; Szeged; Dóra Csendes

Iceland: Institute of Linguistics, University of Iceland; Reykjavík; Eiríkur Rögnvaldsson

Icelandic Centre for Language Technology; Reykjavík; Eiríkur Rögnvaldsson

Ireland: National University of Ireland; Galway; Sean Ryder

Israel: Technion-Israel Institute of Technology; Haifa; Alon Itai

Italy: Dipartimento di Linguistica Teorica e Applicata, Università di Pavia; Pavia; Andrea Sansò

Istituto di Linguistica Computazionale; Pisa; Nicoletta Calzolari

Department of Computer Science, University of Rome "Tor Vergata" ; Rome; Fabio Massimo Zanzotto

European Academy Bozen/Bolzano; Bolzano; Andrea Abel

Latvia: Institute of Mathematics and Computer Science, University of Latvia; Riga; Inguna Skadina

Tilde; Riga; Inguna Skadina

Lithuania: Institute of the Lithuanian Language; Vilnius; Daiva Vaisniene

Center of Computational Linguistics, Vytautas Magnus University ; Kaunas; Ruta Marcinkeviciene

Luxembourg: European Language Resources Association (ELRA); Luxembourg; Bente Maegaard

Malta: University of Malta, Dept. of computer science; Malta; Michael Rosner

Netherlands: Meertens Institute; Amsterdam; H.J. Bennis

Data Archiving and Networked Services; Den Haag; Henk Harmsen

University of Twente, Human Media Interaction Group; Enschede; Roeland Ordelman

Center for Language and Cognition; Groningen; Wyke van der Meer

Digital Library for Dutch Literature; Leiden; C.A. Klapwijk

Instituut voor Nederlandse Lexicologie; Leiden; Remco van Veenendaal

Leiden University Centre for Linguistics; Leiden; Jeroen van de Weijer

Centre for Language Studies, Radboud University; Nijmegen; Pieter Muysken

Centre for Language and Speech Technology, Radboud University; Nijmegen; L. Boves / N. Oostdijk

Max-Planck-Institute for Psycholinguistics; Nijmegen; Peter Wittenburg

University of Utrecht/Netherlands Graduate School of Linguistics; Utrecht; Jan Odijk

ILK Research Group ; Tilburg; Antal van den Bosch

Huygens Instituut KNAW ; Den Haag; Karina van Dalen-Oskam

Norway: Dept. of Culture, Language and Information Technology; Bergen; Koenraad de Smedt

Department of Linguistics and Nordic Studies, University of Oslo; Oslo; Janne Bondi Johannessen

Det humanistiske fakultet, Universitetet i Tromsø; Tromsø; Trond Trosterud

Norwegian University of Science and Technology; Trondheim; Torbjørn Nordgård

Poland: University of Wroclaw ; Wroclaw; Adam Pawlowski

Institute of Applied Informatics, Wroclaw University of Technology; Wroclaw; Maciej Piasecki

Institute of Computer Science, Polish Academy of Sciences ; Warsaw; Adam Przepiórkowski

Institute of English Language, Univeristy of Lodz; Lodz; Lukasz Drozdz

Institute of Slavic Studies, Polish Academy of Sciences ; Warsaw; Violetta Koseska-Toszewa

Portugal: University of Lisbon, NLX-Natural Language and Speech Group; Lisbon; António Branco

Romania: Al.I.Cuza; Iasi; Dan Cristea

Institute for Computer Science, Romanian Academy of Sciences; Iasi; Horia-Nicolai Teodorescu

Research Institute for Artificial Intelligence, Romanian Academy of Sciences; Bucharest; Dan Tufis

University Babes-Bolyai; Cluj-Napoca; Doina Tatar

Serbia: Faculty of Mathematics, University of Belgrade; Belgrade; Duško Vitas

Slovenia: Josef Stefan Institute; Ljubljana; Tomaž Erjavec

Alpineon d.o.o. ; Ljubljana; Jerneja Žganec Gros

Spain: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra; Barcelona; Núria Bel

Universitat de Lleida ; Lleida; Gloria Vázquez

TALG Research Group (University of Vigo) ; Vigo; Xavier Gómez Guinovart

Sweden: Lund University; Lund; Sven Strömqvist

Språkbanken, Dept. of Swedish Language, Göteborg University; Gothenburg; Lars Borin

Dept. Speech, Music and Hearing, CSC, KTH ; Stockholm; Kjell Elenius

Uppsala University, Department of Linguistics and Philosophy; Uppsala; Joakim Nivre

Department of Linguistics; Göteborg; Anders Eriksson

Department of Computer and Information Sciences, Linköping University; Linköping; Lars Ahrenberg

Swedish Institute of Computer Science AB ; Stockholm; Björn Gambäck

Language council of Sweden ; Stockholm; Rickard Domeij

HUMlab, Umeå University ; Umeå; Patrik Svensson

Turkey: Sabanci University – Human Language and Speech Laboratory; Istanbul; Kemal Oflazer

UK: Department of Linguistics and English Langauge, Lancaster University; Lancaster; Anna Siewierska

Oxford Text Archive; Oxford; Martin Wynne

University of Sheffield; Sheffield; Wim Peters

University of Surrey; Guildford; Lee Gillam

Research Institute of Information and Language Processing at the University of Wolverhampton ; Wolverhampton; Gina Sutherland

Language Technologies Unit, Bangor University; Bangor; Briony Williams

Department of English, The University of Birmingham; Birmingham; Oliver Mason

CLARIN newsletter • *publisher* CLARIN • *editor-in-chief* Marko Tadić • *editorial board* Gerhard Budin, Dan Cristea, Koenraad De Smedt, Laurent Romary, Martin Wynne • http://www.clarin.eu